# ALLEVIATING THE CLASS IMBALANCE PROBLEM USING DATA LEVEL APPROACH IN NOISY IMBALANCED DATA-SETS

## Protocol for PhD thesis


**Supervisor**

Dr. Bindiya Ahuja

Professor
Department of computer science
Lingaya's Vidyapeeth
Faridabad, Haryana

**Candidate**

Kamlesh Upadhyay
Roll no - 19PHCS02 (W)


**Department of Computer Science**

**Lingaya's Vidyapeeth, Faridabad, Haryana**

**January 2023**

***Abstract***

Classification is used for to detect different kinds of patterns from the data-set. Although, the classification techniques are very much successful in solving real life problems but are not so successful with unbalanced data-sets. Considering the real time situations, sometimes we are interested to detect the exceptional cases like credit card frauds, fraudulent telephone calls, shuttle system failure, text classification, nuclear explosions, oil spill detection, detection of brain tumor images etc. If we use traditional classification techniques (TCT) in such scenarios, it fail to detect smaller class. Such problem is known as class imbalance problem (CIP). Researchers have proposed many techniques in various diversified directions to address this issue. Current paper is a survey of Data Level binary classification techniques to solve CIP. The focus of the paper is to provide critical review of the topic, state of the art of the data level binary classification techniques developed in this field, details of various metrics used to estimate and identify the performance of various methods used to tackle CIP. For the development of the research area, research directions are also given in the paper so that the research in this field can be extended and in future more enhanced solution towards this problem can be provided. Conventional classification algorithms do not provide accurate results when the data distribution (class sizes) is unequal or data is corrupted with noise because the results are biased towards the bigger class. In many real life cases, there is a requirement to uncover unusual/smaller classes. There are a bundle of examples where importance of smaller/rare class is much-much higher than the bigger class for example- brain tumor detection, credit card fraud or anomaly detection and many more. This is usually called as problem of imbalance classes. The situation becomes worst when the data is corrupted with extra impurities like noise in data or overlapping of class or any other glitch in data because in this scenario traditional methods produce more poor results. This paper proposed a fast, simple and effective data level hybrid technique based on fuzzy concept to overcome the class imbalance problem in noisy condition. To appraise the classification performance of the offered technique it is tested with 40 UCI real imbalanced data sets having imbalance ratio ranges from 1.82 to 129.44 and compared with 12 other approaches. The outcome

specifies that the presented hybrid data level technique performed better and in a fast manner when compared to other approaches.

Data distribution among the classes plays important role during classification process. Traditional classification techniques (TCT) work efficiently with the balanced data or the data with similar size of classes. Their internal design favors the balanced data-sets. TCT fail to detect classes with unbalanced data-set, because the internal design of the algorithm deviates the results towards the bigger class. In some cases TCT ignore the smaller class as noise.

***Keywords— Class Imbalance Problem, Binary Classification, Data Level Approaches, Oversampling, Undersampling, Data level approaches; fuzzy concept; imbalanced data-sets; classification***

## Introduction

### What is DATA MINING

**It** is a process of analyzing and uncovering a meaning full outlines in the large data sets which is known as Big Data [3]. It is a section of computer science which combines many techniques from statistics, data science, database theory and machine learning. Data mining programs break down patterns and connections in data based on what information users request or provide

Background of Data Mining

A number of researches are being done about the data mining [1,2] because data is the key point of any technique. Even though the technique is small or large, old or even new but without data neither we cannot define our problem nor our solution statement.

In 1763, Thomas Bayes explained a theory for relating current to prior probability called the Bayes' theorem. This is the basic of data-mining and probability, since based on estimated probabilities it allows understanding of complex realities [15]. In **1805**, Adrien-Marie Legendre and Carl Friedrich Gauss apply regression to determine

the orbits of bodies about the Sun (comets and planets). The goal of regression analysis is to estimate the relationships among variables. Regression is one of the key tools in data mining. In 1936, Alan Turing introduced the idea of a Universal Machine [5] capable of performing computations like our modern day computers. The modern-day computer is built on the concepts pioneered by Turing.

The word "data mining" appeared in the database community. Retail companies and the financial community are using data mining to analyze data and recognize trends to increase their customer base, predict fluctuations in interest rates, stock prices, and customer demand.

Data mining, *the extraction of hidden predictive information from large databases*, is a powerful technology with great potential to help companies focus on the most important information in their data warehouses. It converts piles of data into useful knowledge that can help the data owners/users make informed choices and take smart actions for their own benefit. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining [4] move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss

Classification Techniques

From all the above-mentioned data mining techniques, classification is the most commonly used technique [8] which contains a set of pre classified samples to create a model which can classify the large set of data. This technique helps in deriving important information about data and metadata (data about data). This technique is closely related to cluster analysis technique, and it uses decision tree or neural network system. There are two main processes involved in this technique.

- Learning – In this process the data are analyzed by classification algorithm.
- Classification – In this process the data is used to measure the precision of the classification rules.

In the research literature, so many classification techniques are proposed to solve real

life problems like Decision Tree, SVM, Neural networks [37] etc.

Class Imbalance Problem

Data distribution among the classes plays important role during classification process. Traditional classification techniques work efficiently in classifying the data having symmetric distribution i.e. the number of instances in the classes are approximately equal. Their internal design favors the balanced data-sets. Traditional Classification Algorithms [17] fail to detect classes when used with unbalanced data-set, because as per their internal design the results in case of unequal size of classes deviates towards the bigger class. These algorithms ignore the smaller class as noise.

In real life situations, sometimes there is a need to detect exceptional cases e.g. credit card frauds, tumor detection, fraudulent telephone calls, Shuttle system failure, text classification, oil spill detection, web spam detection, risk management, information retrieval, intrusion detection, earthquake and nuclear explosion, helicopter gear-box fault monitoring etc. In such cases, Traditional Classification [5] Algorithms do not work well. This problem is identified as Class Imbalance Problem. Class Imbalance problem is the classification problem wherein we are using traditional classification algorithms to classify data with unequal size classes and our objective is to identify smaller class from the data. Researchers have addressed this problem in various diversified ways and a new field of research has emerged under the name Class Imbalance Learning (CIL) and it is evolving day by day. In many papers it is referred as dealing with IDS (Imbalanced data sets) or with rare cases or dealing with skewed data sets (SDS) or skewed distributions. The smaller class in CIL is known as minority class and bigger class is known as majority class.

Class Imbalance Problem does not exist if the purpose is to identify majority class, it actually exists because the purpose is to identify minority class. The ratio of the number of majority to minority class data instances is called imbalance ratio. The problem becomes more risky as this ratio increases i.e. when data-set is highly unbalanced. The techniques proposed by researchers to solve the Class Imbalance Problem are majorly classified into data-level approaches (Pre-processing techniques), algorithm level approached and their hybrid forms. In data-level approach, the researchers have tried to balance the data-sets before applying traditional classification algorithms so that results may not overwhelmed by the majority class [39]. In algorithm level approaches,

the researchers have worked upon the internal algorithm structure and tried to work upon the sensitivity of algorithm towards the majority class. These algorithms comes under the category of cost sensitive algorithms. Third approach is the hybrid form, which is the combination of data-level and algorithm level approaches [17, 19]. Latest concept used in creating class imbalance techniques is the ensemble approach, which uses bagging and boosting methods to enhance the traditional classifiers.

## Review of literature

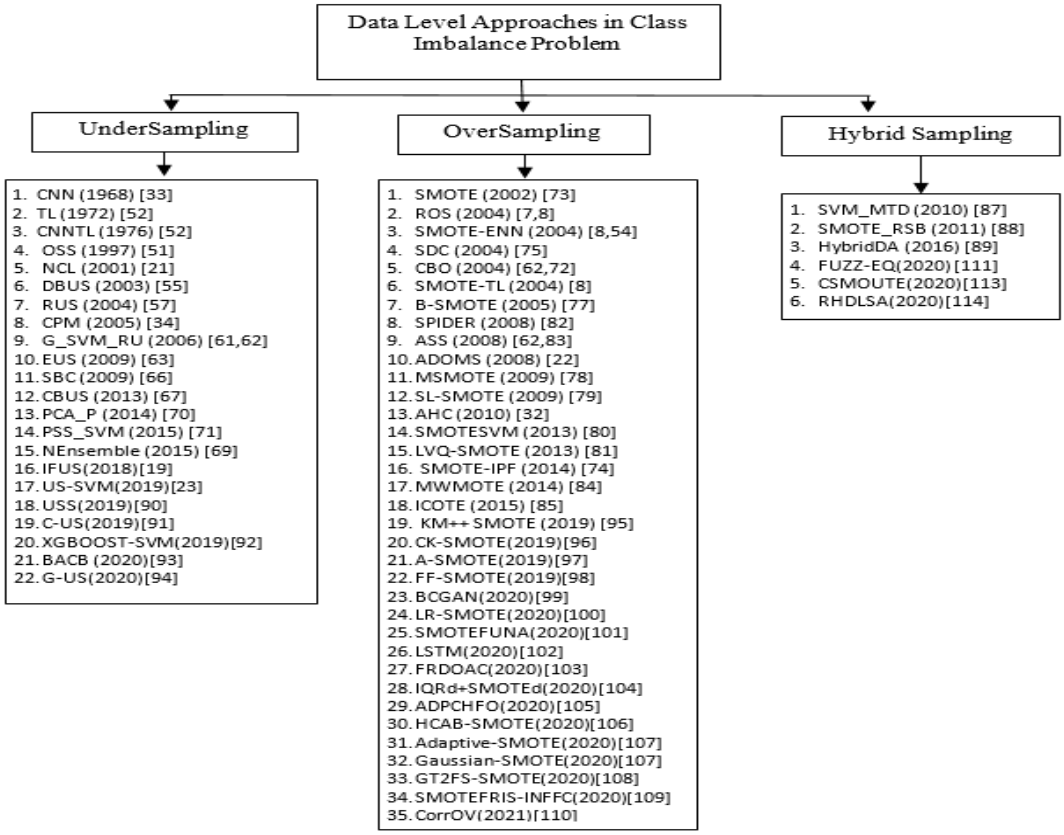
In real life situations, sometimes there is a need to detect exceptional cases e.g. fraud over credit card, tumor detection, fraudulent telephone calls, Shuttle system failure, text classification, detection of oil spill, web-spam detection, risk management, information retrieval, intrusion detection, earthquake and nuclear explosion, helicopter gear-box fault monitoring [12] etc. In such cases, TCT do not work well as they are unable to detect smaller class. This problem well known as Class Imbalance Problem (CIP). Researchers have addressed this problem in various diversified ways and a novel area of research is emerged for the name Class Imbalance Learning which is evolving day by day. In many papers it is referred as dealing with IDS (Imbalanced data sets) or with rare cases or dealing with skewed data sets or skewed distributions. Many researchers have written about the work done in this field. [2-18] In 2015, Ali et al. done a survey on Class imbalance problem. He discussed various issues, problems with class imbalance problems and done a survey of classifiers designed to handle CIP till 2015. A recent survey, Haixiang et al. discussed the technical as well as the practical aspects of the class imbalance problem with some future directions.[18] This paper has covered in details the data level approaches of the class imbalance problem till 2021. The survey is divided into three key categories of Data Level Approaches as Undersampling, Oversampling and Hybrid-data-level approaches.

The rest of survey is organized in following manner. Section 2 covers the state of the art regarding data level binary classifiers developed to address CIP year wise. Section 3 discusses various performance metrics developed in the class-imbalance domain for evaluate the classifier's performance. In the 4 section, suggestions are given which may

help the researchers to extend the work going on in this field to solve CIP in a better manner.

State of the Art on Data Level Binary Classifiers to handle Class Imbalance Problem

Researchers have been proposed various solutions to deal with CIP in diversified directions. We have divided this survey into 3 categories, which are based on the concepts and approaches used in the techniques as shown in Figure 1. Before exploring class imbalance learning techniques, we want to share some terminology of this field referred in various papers. The smaller class in class imbalance learning is known as minority (MI) class and bigger class is known as majority (MA) class. CIP does not exist if the purpose is to identify MA class, it actually exists because the purpose is to identify MI class. The ratio of the number of MA to MI class data instances is called imbalance data-set is highly unbalanced. The techniques proposed by researchers to solve the CIP are mainly/key classified to data(level) approaches-090 (Pre-processing techniques), algorithm level approached and their hybrid forms.



The problem becomes more risky as this ratio increases i.e. when data-set is highly unbalanced. In data-level approach, the researchers have tried to balance the data-sets before applying traditional classification algorithms so that results may not overwhelmed by the MA class. In algorithm level approaches, the researchers have worked upon the internal algorithm structure and tried to work upon the sensitivity of algorithm towards the MA. These algorithms comes under the category of cost sensitive algorithms.[24-38] Third approach is the hybrid form, which is the combination of data-level and algorithm level approaches [35] This section explores various data level techniques developed to alleviate Class imbalance problem.

*Data Level Techniques*

Data-level-techniques are those wherein the data is pre-processed, before classifying the data, using traditional classifiers/models. The main goal of the data level techniques is to rebalance the data before classification. The root cause of the CIP is imbalanced data-sets i.e. one class has more number of instances than other and the result of smaller class is influenced by the bigger class. [8] Moreover, the risk factor is directly proportional to the imbalance ratio i.e. as the data-set become more imbalance the impact of MA class is more on the results of MI class. In some of the papers, these techniques are referred as external techniques. Data level techniques are further classified as undersampling, oversampling and hybrid approach. Following sub-section explores various algorithms proposed by the researchers under data-level category.

*Algorithms based upon Undersampling*

In the Undersampling approach, the working area in the data set is MA class wherein the instances are removed from the MA class either randomly or by using some technique to improve the imbalance ratio before classifying the data. Major advantage of undersampling approach over oversampling is that it is fast and simple. In case of one-class identification problem, undersampling is better over oversampling & hybrid approach as we are removing data from MA class, which is not the region of interest. But in case of binary or multiclass detection, it may loss of informational data contained in the bigger class. Various techniques studied under this category are as below:

*Condensed Nearest Neighbor (CNN), 1968:* Hart's Condensed-Nearest-Neighbor Rule (CNN)[33] by which we find a steady subset of points. Its goal is to decrease the size to actual design by the removal of certain points without any affecting meaningfully the

outcome of NN classification. In this technique, all the data-points are separated into three categories: outliers, prototypes and absorbed points. Outliers are those points which does not follow the minimum requirement of points in its neighborhood. Prototype specifies the least requirement of points in the neighborhood. Absorbed points are those which are not outliers and can be accepted as good points. The algorithm checked all the points one by one and assign the points according to these conditions.

*Tomek Links (TL), 1972:* Tomek links [34] are used to eradicate the data points hence can be used as an undersampling method wherein the points from the majority class are eliminated. Tomek links can be well-defined as follows:

Let we have two points 'a' and 'b' belonging to dissimilar classes, and d(a,b ) is distance of 'a' - 'b'. A (a,b) together is called a Tomek link if no other point 'c' exists such that d(a,b) < d(a, c ) or d(a,b) < d(b,c). If the pair forms a Tomek link, and then one point would be noise or both points are borderline.

*One sided selection (OSS), 1997:* This method is the application of Ordered Condensed Nearest Neighbor OCNN by first selecting the borderline instances of MA class and then using Condensed Nearest Neighbor (CNN) technique to remove those instances [7-10]. Borderline instances can be taken as 'unsafe instances' as a minor amount of noise may root them to drop on the incorrect side of class. OCNN method[7, 52] is the modification of CNN[53] method which randomly remove instances based upon the nearest neighbor rule. This method tries to remove the borderline instances. It is based upon the logic that if x and y are two instances belong to two different classes and d(x,y) is the minimum distance without any other instance comes in-between. Then either one of these instances is a noise or both are considered as the borderline instances of the classes, which can further removed from the data set using CNN rule. OCNN and CNN are using nearest neighbor rule (NNR), proposed by Wilson in 1972. In this method Wilson proposed a process, wherein the data instances are reduced in a pre-classified instance using nearest principle. It uses 3-nearest-neighbor rule and then classification by single-nearest-neighbor rule with remaining pre-classified instances. After pre-processing data, OSS method uses C4.5 to classify it.

*Condensed Nearest Neighbor + Tomek Link (CNNTL), 1976:* It is similar to OSS and also uses the concept of CNN and TL to reduce the data points from the majority class [25] It is different from OSS in the sense that it implement CNN first and then is followed by applying Tomek Link on the data set. Tomek links are computationally

expensive so to reduce the computational time Tomek links are applied after CNN on the reduced data-set.

*Neighborhood Cleaning Rule (NCL), 2001:* Neighborhood Cleaning Rule (NCL) [21] uses the Wilson's Edited Nearest-Neighbor (ENN) [10] to eliminate majority-class data. ENN eliminates any point whose class label varies from class of minimum two out of three nearest neighbors. NCL adjusts the ENN and increase the chance of data-cleaning. algorithm can be defined in the following manner for a two-class problem: for each point 'a' in the training data set, that is three nearest neighbors are found. If 'a' goes to majority-class instance and the classification given by its 3-nearest-neighbors contradicts the original class of 'a', then 'a' is removed. If 'a' goes to the minority-class and its 3 nearest neighbors misclassify 'a', then nearest neighbors that belong to the higher class are deleted.

*Distance Based Undersampling (DBUS), 2003:* This undersampling approach was given by Y. M. Chyi. In this method the instances from MA class are removed based upon four different distance measures. These are nearest distance, farthest distance, average nearest distance and average farthest distance. In the nearest distance approach, the distance of all MA instances are calculated from MI class and those instances which are nearest to the MI class are removed. In the farthest distance approach, the distance of instances in the MA class are calculated and those which are at the maximum distance from MA class are removed. In the third and fourth approach, the average nearest and average farthest distance is calculated from MI and MA classes. Considering the large dataset, the major limitation of this method is that it spends a lot of time in selecting the instances from the MA class based upon these distances. Hence this method is not suitable for real and large data-sets. [11-13]

*Random Undersampling (RUS):* This is the simplest undersampling technique by randomly removing instances from the MA class. Although, it does not make any intelligent effort to remove instances but reduces the time needs to build the model which is the key advantage when making an ensemble-models.

*Algorithms based upon Oversampling*

Oversampling is the process of increasing the number of data instances into the MI class either randomly through replication of the existing data or generating synthetically by using any technique to improve the imbalance ratio so that TCT can be used to classify the data. Major advantage of oversampling approach is that no data is lost

during the process as we are not removing any data from it, although new information is generated for rebalancing. The limitation of oversampling is that it is computationally expensive as compared to undersampling approach because we are generating data. It may also face overfitting problem, it will replicate the same data instances to rebalance the data-set. If we consider the three categories (One class or binary or multiclass detection), oversampling changed the MI class data by adding more instances, which in turn changed the actual data in the class, which is not acceptable in some critical real time problems [4-5]. Various techniques studied under this approach are:

*Random Oversampling (ROS), 2004:* In this method the MI instances are increased by randomly replicating the MI instances to balance the original data-set. [14] The disadvantage with the method is that it can rise the possibility of over fitting, as it is replicating the same instances.

*Cluster based oversampling (CBO), 2004:* The technique is proposed to deal with the class imbalance within class. It uses k-means clustering technique to detect random cluster and calculate the cluster center by the mean of feature vector. After that, every cluster is compared with other by finding the Euclidean distance and then allocated to the class having minimum Euclidean distance.

*Synthetic MI Oversampling Technique (SMOTE), 2002:* In this technique [37] data instances are synthetically generated in the MI class rather than replacement. In this method, oversampling is done by considering all the MI class data and new data is generated by using interpolation method. Based upon the amount of oversampling required, neighbors from the 'k'-nearest neighbors are chosen randomly. SMOTE has considered five nearest neighbors. One of the main limitation of SMOTE is overgeneralization as SMOTE blindly generalizes data instances in the MI class without checking the status of instances as good instances or noise.[74] In case of SMOTE, if the classes are not well defined and there is a possibility of overlapping between them, the interpolation may expand the smaller class cluster, which may result into more overlapping between the classes. This situation can cause overfitting. Up till now, many variations of SMOTE are proposed by the researchers. In 2004, SMOTE-TL is proposed, wherein Tomek links are used to remove the noisy points or the points which are lying on the boundary. Tomek link [7] is defined as a pair of points 'a', 'b' such that there should not exist any other point 'c' so that

$$d(a,b) < d(a,c) or\ d(a,b) < d(b,c) \quad (1)$$

Where ‘d’ is the distance measure. If two of these points form a Tomek link then either one of them is a noise or both are lying on the boundary. In the same year, to remove the limitation of overfitting, SMOTE with Edited Nearest Neighbor (SMOTE-ENN) is proposed *by* Batista et al. by applying Wilson’s ENN rule after SMOTE application to remove any instance misclassified by its three nearest neighbors from the training set. In the same year, another variation of SMOTE, SMOTE with Different Costs (SDC) is proposed by combining SMOTE with Different error cost algorithm. SDC uses the concept of different error cost for dissimilar classes to send the boundary away from the positive data instances and uses SMOTE to increase the strength of smaller class to make the boundary better.

In 2005, Han et al. proposed another intelligent method Borderline-SMOTE (B-SMOTE) to improve the performance of SMOTE by over-instanced only the borderline instances of MI class.[76, 77] The method is based upon the concept that the data instances which are close to or on the borderline can be misclassified with greater than those which are far from the borderline thus are more important for classification.

In 2009, S. Hu et al. proposed a robust technique based upon SMOTE to tackle the noise present in the data. Modified SMOTE (MSMOTE) is the variation of SMOTE, which apart from considering the distribution of MI class also tackles the noisy instances based upon the k-NN classifier method.[9,78] MSMOTE classifies the MI class into three different classes namely security, Border and latent noise instances. If the label of the instance is same as that of the label of all the instances of its k-nearest neighbors then the instance is allocated to security class. If its label is entirely different from the labels of its k-nearest neighbors then it is allocated to the noise class. Any instance, which does not follow the conditions of either security or noise class is assigned to the border class. MSMOTE generates the synthetic instances differently for these three classes. For the security class, it randomly selects the instance from the k-nearest neighbors. For the border class, it selects the nearest neighbor and no instance is selected from the latent noise class. Another variation, Safe-Level SMOTE (SL-SMOTE) is proposed in the same year by Bunkhumpornpat et al., which assigned a safe level to every positive instance before synthetically generating data instances in the MI class [17]. The algorithm is designed in such a way that every synthetically generated instance lie near to the largest safe level so that all the instances which are generated synthetically lie in the safe regions only. Safe levels is defined by k-nearest neighbor. Any instance whose safe level value is closer to zero, is considered as noise.

SMOTE introduces synthetic instances using the interpolation method between two instances of a data-set to increase the size of MI class. Due to blind selection of instances to generate other instances, noise or superfluous instances may be added into the data-set, which may degrades the classifier's performance. To remove this limitation another variation, SMOTE-SVM is proposed in 2013. The technique uses Support vector machines (SVM) in the learning framework of SMOTE to improve the classification accuracy. Instead of using traditional classifiers like decision tree etc. it uses the power of SVM after balancing the data-set through SMOTE technique. As per the technique, the data-set is divided into number of sub-sets and then from every sub-set, MI and MA classes are identified. After enhancing the instances in the MI class using SMOTE, the sub-set is again joined and SVM is implemented to classify the data. In the same year, another variant of SMOTE based on learning vector quantization (LVQ-SMOTE) is proposed by Nakamura et al.[81] This technique generates codebooks of data using LVQ technique. Each generated instance is compared with codebook of reference data. The instance is selected if the codebook in target dataset is similar to reference dataset.

Another variant of SMOTE, SMOTE-IPF is proposed in 2014. It extended SMOTE with a noise filter called Iterative-Partitioning Filter (IPF), which iteratively removes noisy instances present in the data-set. IPF removes noisy instances in multiple iterations.[74] The iteration process stops when the total known noisy instances in every iteration becomes less than a specified percentage of the size of the original data-set. In the year 2019, Chen et al. proposed an improved version of SMOTE (KM++SMOTE) which handled the problem of new synthetic sample distribution marginalization and uneven distribution of overall data set caused by SMOTE.[95] In the same year, Guo et al. proposed another SMOTE based approach (CK-SMOTE), which used canopy and K-means method to cluster the data before using SMOTE. [18]Advanced SMOTE (A-SMOTE) was proposed by Hussein et al. in the same year to adjust the new added instances of MI class. This method first apply SMOTE to generate synthetic data within MI class and then remove those points that are more nearer to the MA class than MI. In the same year Kaur and Gosain proposed firefly based SMOTE (FF-SMOTE) which used firefly concept to generate the synthetic data rather than interpolation concept within MI class. In 2020, another variation of SMOTE is proposed (IQRd+SMOTEd) which identifies outliers using interquartile range algorithm and then apply SMOTE to balance the data. Another variation of SMOTE, Hybrid Clustered Affinitive Borderline

SMOTE (HCAB-SMOTE) is proposed in the same year by Majzoub et al. [21]to tackle the generation of less useful data. This method removes the majority noise instances and enhance the density of borderline using oversampling.[106]

*Generative Adversarial networks based OS (BCGAN), 2020:* This method generates instances within MI class based on borderline-conditional generative adversarial networks to balance the data before classification. [20]

*Long Short Term Memory based OS (LSTM), 2020:* This method used Long short term memory depending upon deep-learning method to calculate association information from the manifolds to MI class and produces new MI sequence-based on the conditional possibility on the parent orders to generate the synthetic-data. [10]

*Fuzzy Representativeness Difference based OS (FRDOAC), 2020:* This method proposed a fuzzy representativeness difference based oversampling technique, which uses affinity propagation and chromosome theory of inheritance to generate synthetic samples.

*Adaptive Weighted OS (ADPCHFO), 2020:* This method proposed an adaptive weighted oversampling approach, which is based on density peal clustering with heuristic filtering. It applies modified density peak clustering to cluster the MI class due to its capability of accurately identifying sub-clusters with different sizes and densities to tackle the 'between the class' and 'within class'-imbalance issues.

Co-relation based OS (CorrOV), 2021: This paper proposed co-relation based oversampling approach with cost-sensitive ensemble learning to classify the data [11]

*Hybrid Sampling*

All those techniques that uses both oversampling as well as undersampling approach to improve the imbalance ratio (IR) before classification process comes under the HS approach.

*SVM_MTD, 2010:* This technique oversampling the MI class and undersampling the MA class to balance the data-set. It uses Gaussian Fuzzy membership function and α-cut for oversampling and uses mega-trend Diffusion membership[87] function to generate new instances in the MI class. To enhance the classification accuracy, it transforms the data attribute dimensions from 2-dimentional space to higher dimension space by means of classification-based information.

*SMOTE_RSB, 2011:* This study proposed new hybrid approach by combining undersampling and oversampling. It combined SMOTE with a technique based upon rough sets. In this technique the quality of new synthetic data generated is controlled by using rough set theory.

*HybridDA, 2016:* Akbani and al-Rifaie proposed hybrid technique by using SMOTE for oversampling, RUS for undersampling and SVM for classification. They used grid search for selecting the optimized values of misclassification cost, gamma and kernel type. [12]

*Probability Integral Transform based US (FUZZ-EQ), 2020*: This method adjusts the training data to a imaginary uniform partition before classification. The actual data distribution transformed to a uniform distribution by applying the probability integral transform.

*Synthetic concept based Hybrid sampling (CSMOUTE), 2020:* This method combined Undersampling and Oversampling approach based upon synthetic generation and removal of instances from MI and MA class respectively.

*Robust Hybrid Data Level Sampling (RHDLSA), 2020:* This paper proposed a hybrid sampling approach to handle multiple data impurities like class imbalance, noise and dissimilar data distribution among the classes.

**History and current situation**

Researchers have developed many techniques to resolve this problem. Majorly the problem is tackled at three levels:

- Data Level Approach
  - UnderSampling [3]
    - Instances in the majority class are decreased
  - OverSampling [4]
    - Instances in the minority class are increased
  - Hybrid Sampling [8]
    - Combination of Undersampling and Oversampling
- Algorithm Level Approach
  - Cost Function based Approach [6]

- More weights are assigned for the negative instances or vice-versa

- Hybrid Level Approach (Ensemble Learning Method)
  - Combination of Data level and Algorithm level Techniques [7]

Below are some examples or existing techniques.

Undersampling Methods

*Distance Based Undersampling (DBUS), 2003:* This undersampling approach was given by Y. M. Chyi. (Chyi 2003) In this method the instances from MA class are removed based upon four different distance measures. These are nearest distance, farthest distance, average nearest distance and average farthest distance. In the nearest distance approach, the distance of all MA instances are calculated from MI class and those instances which are nearest to the MI class are removed. In the farthest distance approach, the distance of instances in the MA class is calculated and those which are at the maximum distance from MA class are removed. In the third and fourth approach, the average nearest and average farthest distance is calculated from MI and MA classes. Considering the large dataset, the major limitation of this method is that it spends a lot of time in selecting the instances from the MA class based upon these distances. Hence this method is not suitable for real and large data-sets (Yen and Lee 2009).

*Random Undersampling (RUS):* This is the simplest undersampling technique by randomly removing instances from the MA class [14-15] Although, it does not make any intelligent effort to remove instances but reduces the time needs to build the model which is the key advantage when making an ensemble-models.

*Evolutionary Undersampling (EUS), 2009:* This undersampling method uses the principle of evolutionary prototype selection for imbalanced data. In EUS, each chromosome is represented by a binary vector, showing the presence and absence of the instances in data-set (García and Herrera 2009) EUS uses 1NN classifier and geometric mean to estimate the performance. It makes use of a well-known CHC algorithm [23] to evolve the initial random population. CHC is the genetic algorithm that uses heterogeneous uniform crossover (HUX) to combine two chromosomes. In case of EUS, the HUX of the original CHC algorithm is modified so that good reduction rate is obtained and hence reducing the chance of including instances in a date-set. In

this way each time HUX switches the gene on, the gene is switched off with the sure probability.

*Undersampling with Support Vector (US-SVM), 2019:* This method reduces the size of MA class based upon support vectors that help to engender decision boundary (Arafat and Hoque 2008).

*Undersampling using Sensitivity (USS), 2019:* This paper proposed a novel undersampling method based upon the sensitivity of each MA instance. Noisy and safe instances have low sensitivity whereas the borderline instances have high sensitivity. USS remove points from MA class with high sensitivity [19].

*Ensemble Learning based US (XGBoost-SVM), 2019:* This paper proposed an ensemble-learning based undersampling technique using Extreme Gradient Boosting and SVM.

*Granular Computing based US (G-US), 2020:* The concept used in this method is based upon granular computing. It builds information granules around the selected pattern of MA class and resultant granules to be removed from MA class are selected based upon the highest specificity value.

*Algorithms based upon Oversampling*

Oversampling is the process of increasing the number of data instances into the MI class either randomly through replication of the existing data or generating synthetically by using any technique to improve the imbalance ratio so that TCT can be used to classify the data. Major advantage of oversampling approach is that no data is lost during the process as we are not removing any data from it, although new information is generated for rebalancing. The limitation of oversampling is that it is computationally expensive[31] as compared to undersampling approach because we are generating data. It may also face overfitting problem, it will replicate the same data instances to rebalance the data-set. If we consider the three categories (One class or binary or multiclass detection), oversampling changed the MI class data by adding more instances, which in turn changed the actual data in the class, which is not acceptable in some critical real time problems. Various techniques studied under this approach are:

*Random Oversampling (ROS), 2004:* In this method the MI instances are increased by randomly replicating the MI instances to balance the original data-set. The disadvantage

with the method is that it can rise the possibility of over fitting, as it is replicating the same instances.

*Cluster based oversampling (CBO), 2004:* The technique is proposed to deal with the class imbalance within class [18]. It uses k-means clustering technique to detect random cluster and calculate the cluster center by the mean of feature vector. After that, every cluster is compared with other by finding the Euclidean distance and then allocated to the class having minimum Euclidean distance.

*Synthetic MI Oversampling Technique (SMOTE), 2002:* In this technique, *Chawla and Bowyer* explained that data instances are synthetically generated in the MI class rather than replacement. In this method, oversampling is done by considering all the MI class data and new data is generated by using interpolation method. Based upon the amount of oversampling required, neighbors from the 'k'-nearest neighbors are chosen randomly. SMOTE has considered five nearest neighbors. One of the main limitation of SMOTE is overgeneralization as SMOTE blindly generalizes data instances in the MI class without checking the status of instances as good instances or noise. In case of SMOTE, if the classes are not well defined and there is a possibility of overlapping between them, the interpolation may expand the smaller class cluster, which may result into more overlapping between the classes. This situation can cause overfitting. Up till now, many variations of SMOTE are proposed by the researchers. For example

In 2004, SMOTE-TL is proposed, wherein Tomek links are used to remove the noisy points or the points which are lying on the boundary. Tomek link [25-26] is defined as a pair of points 'a', 'b' such that there should not exist any other point 'c' so that

$$d(a,b) < d(a,c) or\ d(a,b) < d(b,c) \quad (1)$$

Where 'd' is the distance measure. If two of these points form a Tomek link then either one of them is a noise or both are lying on the boundary. In the same year, to remove the limitation of overfitting, SMOTE with Edited Nearest Neighbor (SMOTE-ENN) is proposed by Batista et al. by applying [15,16] after SMOTE application to remove any instance misclassified by its three nearest neighbors from the training set. In the same year, another variation of SMOTE, SMOTE with Different Costs (SDC) is proposed by combining SMOTE with Different error cost algorithm SDC uses the concept of different error cost for dissimilar classes to send the boundary away from the positive data instances and uses SMOTE to increase the strength of smaller class to make the boundary better.

SMOTE introduces synthetic instances using the interpolation method between two instances of a data-set to increase the size of MI class. Due to blind selection of instances to generate other instances, noise or superfluous instances may be added into the data-set, which may degrades the classifier's performance. To remove this limitation another variation, SMOTE-SVM is proposed in 2013. The technique uses Support vector machines (SVM) in the learning framework of SMOTE to improve the classification accuracy [12] Instead of using traditional classifiers like decision tree etc. it uses the power of SVM after balancing the data-set through SMOTE technique. As per the technique, the data-set is divided into number of sub-sets and then from every sub-set, MI and MA classes are identified. After enhancing the instances in the MI class using SMOTE, the sub-set is again joined and SVM is implemented to classify the data. In the same year, another variant of SMOTE based on learning vector quantization (LVQ-SMOTE) is proposed by Nakamura et al [21 , 23]. This technique generates codebooks of data using LVQ technique. Each generated instance is compared with codebook of reference data. The instance is selected if the codebook in target dataset is similar to reference dataset.

Another variant of SMOTE, SMOTE-IPF is proposed in 2014. It extended SMOTE with a noise filter called Iterative-Partitioning Filter (IPF), which iteratively removes noisy instances present in the data-set. IPF removes noisy instances in multiple iterations (Sáez and Luengo 2014) The iteration process stops when the total known noisy instances in every iteration becomes less than a specified percentage of the size of the original data-set.

In the year 2019, Chen et al. proposed an improved version of SMOTE (KM++SMOTE) which handled the problem of new synthetic sample distribution marginalization and uneven distribution of overall data set caused by SMOTE (Chen and Dong 2019) In the same year, Guo et al. proposed another SMOTE based approach (CK-SMOTE), which used canopy and K-means method to cluster the data before using SMOTE [42,44] Advanced SMOTE (A-SMOTE) was proposed by Hussein et al. in the same year to adjust the new added instances of MI class. This method first apply SMOTE to generate synthetic data within MI class and then remove those points that are more nearer to the MA class than MI ( In the same year Kaur and Gosain proposed firefly based SMOTE (FF-SMOTE) which used firefly concept to generate the synthetic data rather than interpolation concept within MI class (Zhang and Wang 2019).

In 2020, another variation of SMOTE is proposed (IQRd+SMOTEd) which identifies outliers using interquartile range algorithm and then apply SMOTE to balance the data (Nnamoko and Korkontzelos 2020). Another variation of SMOTE, Hybrid Clustered Affinitive Borderline SMOTE (HCAB-SMOTE) is proposed in the same year by Majzoub et al. to tackle the generation of less useful data. This method removes the majority noise instances and enhance the density of borderline using oversampling [2,3] In the same year, General Type-2 Fuzzy based SMOTE is proposed (GT2FS-SMOTE), which introduced the synthetic data within MI class by selecting those points as candidate points which are close to the center of the class hence able to avoid selecting outlier points from the MI class [92-94] Pan et al. in the same year proposed two another variations of SMOTE to balance the data. Adaptive-SMOTE modified the SMOTE process by adaptively selecting the groups of inner and danger data from the MI class as the candidate points [15,16]. Gaussian-SMOTE improves the process by combining the dimension reduction with the Gaussian distribution to make the tail of Gaussian distribution thinner to overcome the impact of noise and borderline instances, Bashir et al. in the same year proposed another variation of SMOTE (SMOTEFRIS-INFFC), which uses the concept of Fuzzy rough instance selection type-II and Iterative Noise Filter based on the fusion of classifiers [35].

*Adaptive Synthetic Sampling (ASS), 2008:* He et al. proposed a very simple and effective method to generate synthetic instances. It uses K-nearest neighbor method to select the dense region in the dataset and generate instances within the dense region (He and Garcia 2009, Amin and Answer 2016).

*MA weighted MI oversampling (MWMOTE), 2014:* S. Barua et al. proposed an intelligent technique to generate synthetic instances. The idea is to assign a weight to the hard to learn MI class instances as per their Euclidean distance from the nearest MA class instance. The technique ensures the generation of synthetic instance within MI class (Amin and Answer 2016).

*Immune centroid Oversampling (ICOTE), 2015:* X. Ai et al. proposed a novel technique based upon immune networks. It generates a set of immune centroids to broaden the decision area of the MI class. These centroids act as the synthetic data then using ENN the MA data instances are disposed into MI space (Ai, Wu, Sheng, Zhao, and Cui 2015).

*Generative Adversarial networks based OS (BCGAN), 2020:* This method generates instances within MI class based on borderline-conditional generative adversarial networks to balance the data before classification (Jung, Son and Moon 2020).

*Long Short Term Memory based OS (LSTM), 2020:* This method used Long short term memory depending upon deep-learning method to calculate association information from the manifolds to MI class and produces new MI sequence-based on the conditional possibility on the parent orders to generate the synthetic-data (Pal and Tarafder 2020).
*Fuzzy Representativeness Difference based OS (FRDOAC), 2020:* This method proposed a fuzzy representativeness difference based oversampling technique, which uses affinity propagation and chromosome theory of inheritance to generate synthetic samples (Ren and Young 2020).

*Adaptive Weighted OS (ADPCHFO), 2020:* This method proposed an adaptive weighted oversampling approach, which is based on density peal clustering with heuristic filtering. It applies modified density peak clustering to cluster the MI class due to its capability of accurately identifying sub-clusters with different sizes and densities to tackle the 'between the class' and 'within class'-imbalance issues [17].
Co-relation based OS (CorrOV), 2021: This paper proposed co-relation based oversampling approach with cost-sensitive ensemble learning to classify the data (Devi and Biswas 2020).

*Hybrid Sampling*

All those techniques that uses both oversampling as well as undersampling approach to improve the imbalance ratio (IR) before classification process comes under the HS approach.
*SVM_MTD, 2010:* This technique oversampling the MI class and undersampling the MA class to balance the data-set. It uses Gaussian Fuzzy membership function and α-cut for oversampling and uses mega-trend Diffusion membership (Li and Wn 2007) function to generate new instances in the MI class. To enhance the classification accuracy, it transforms the data attribute dimensions from 2-dimentional space to higher dimension space by means of classification-based information.
*SMOTE_RSB, 2011:* This study proposed new hybrid approach by combining undersampling and oversampling [135,136]. It combined SMOTE with a technique

based upon rough sets. In this technique the quality of new synthetic data generated is controlled by using rough set theory.

*HybridDA, 2016:* Akbani and al-Rifaie proposed hybrid technique by using SMOTE for oversampling, RUS for undersampling and SVM for classification. They used grid search for selecting the optimized values of misclassification cost, gamma and kernel type.

*Probability Integral Transform based US (FUZZ-EQ), 2020*: This method adjusts the training data to a imaginary uniform partition before classification. The actual data distribution transformed to a uniform distribution by applying the probability integral transform [15, 17].

*Synthetic concept based Hybrid sampling (CSMOUTE), 2020:* This method combined Undersampling and Oversampling approach based upon synthetic generation and removal of instances from MI and MA class respectively [10, 12].

*Robust Hybrid Data Level Sampling (RHDLSA), 2020:* This paper proposed a hybrid sampling approach to handle multiple data impurities like class imbalance, noise and dissimilar data distribution among the classes [24].

**Comparison of existing techniques**

This section includes the comparison of different existing techniques in tabular form. Below table includes the description of data-level undersampling techniques.

**TABLE I Description of data- level undersampling techniques**

| **Purposed Technique** | **Tools and Procedure** | **Highlight /Contributions** | **Authors/Published year** |
|---|---|---|---|
| OSS [25] | it is specific with noisy borderline and redundant data, explained (OSS) here one side selection is taken to produce a steady majority class subset post removal the borderline and noisy pattern | Explains noisy and borderline and safe redundant patterns | Kubat and Matwin-1997 |
| IRUS [26] | Its Inverse Random and a Undersampling | Produced high (TRP) during training. | Tahir et al.-2012 |

| | | | |
|---|---|---|---|
| | (IRUS) technique dependent on the previous probability, p of the minority class. And in the order of p2 the Majority class patterns are selected. | | |
| Redundancy-driven-modified Tomek-link [18] | Proposed collective removal of outlier, noisy and redundant patterns and provide elimination and detection for majority class with smallest contribution in valuation of correct class labels. | K-NN classifier performance are improved with it. | Devi et al.-2017 |
| ENN [11] | It Uses ENN rule to identify insignificant pattern of majority class for 3-NN classification rule the ENN rule is explained as the prototype selection method. | Detects the error-prone or error- contributing majority class patterns. | Laurikkala-2019 |
| EUS [27] | in this method majority classes are undersampled multiple times to produced a set of results and then its evaluated through a fitness function. | Accomplishes higher accuracy when there is highly imbalanced datasets | Garcia et al.-2009 |
| K-NN sampling [28] | Working K-NN strategy is used for undersampling technique where each NN is defined by a distance function. Also including RUS. Pattern are noticed based on majority pattern over distance of minority, where K is specific to problem. | K-NN approximation with the minority as well as majority NNs. | Zhang et al.-2003 |

Below table includes the description of various oversampling techniques.

**TABLE II Description of various Oversampling techniques**

| **Purposed Technique** | **Tools and Procedure** | **Highlight /Contributions** | **Authors/Published Year** |
|---|---|---|---|
| ROS [31] | In this method the MI instances are increased randomly by replicating the minority instances to balance the original data-set . The main disadvantage of ROS method is that it can increase the possibility of over fitting, as it is replicating the same instances. | Its first technique to increase the data randomly using oversampling. | A. Fernandez et al-2004 |
| SMOTE[10] | In this technique, data instances are synthetically generated in the minority class rather than replacement. In this method, oversampling is done by considering all the minority class data and new data is generated by using interpolation method | It is more stable that ROS, it is the first technique to generate data synthetically. | N. V. Chawla et al-2002 |
| CBO[32] | Cluster based oversampling is provided to deal with the class imbalance within class. It uses k-means clustering technique to detect random cluster and calculate the cluster center by the mean of feature vector | In this technique the class is divided into cluster and then cluster are compared with other by finding the euclidean distance and then allocated to the class having minimum Euclidean distance | H. He and E. A. Garcia-2004 |
| ADASYN[33] | Adaptive Synthetic Sampling He et al. proposed a very simple and effective method to generate synthetic instances. It uses K-nearest neighbor method to select the dense region in the dataset and generate | It is a synthetic oversampling method, sometime more effective than SMOTE, its used KNN method. | He et al-2008 |

| | | | |
|---|---|---|---|
| | instances within the dense region | | |
| MSMOTE[34] | Modified SMOTE – it consider two thing first distribution of minority class samples, and second eliminates noise samples by excepting mediation. The mixture of MSMOTE and AdaBoost are used at several highly and considerable imbalanced data sets. | The comparable results show that the calculational performance of MSMOTE is far better than SMOTEBoost the minority class and F-values both are improved | Shengguo Hu,Yanfeng Liang,Ying He-2009 |

Below table. includes the description of various Hybrid Data Level techniques.

**TABLE III. Description of various Hybrid Data level Techniques**

| **Purposed Technique** | **Tools and Procedure** | **Highlight /Contributions** | **Authors/Published Year** |
|---|---|---|---|
| SVM_MTD[37] | In this technique majority classes are over instances where as minority classes can be under instance, It uses Gaussian-Fuzzy-membership function and α-cut for oversampling and uses mega-trend-Diffusion-membership function to generate new instances for minority class | To enhance the classification accuracy, it transforms the data attribute dimensions from 2-dimentional space to higher-dimension space by using classification associated information. | Li Der-Chiang, Chiao-Wen Liu and Susan C. Hu-2010 |
| SMOTE_RSB[38] | This study proposed new hybrid approach by combining undersampling and oversampling. It combined SMOTE with a technique based upon rough sets | In this technique the quality of new synthetic data generated is controlled by using rough set theory. | Ramentol Enislay et al-2012 |
| HybridDA[39] | The LRSMOTE algorithm used traditional SMOTE. The described method based on 2 step filtering firstly it uses SVM and k-means to reduce noise in data set that come from original sources, and then changes the formula generated by the new | They used grid search for selecting the optimized values of misclassification cost, gamma and kernel type . | Akbani and al-Rifaie - 2016 |

| | | | |
|---|---|---|---|
| | sample. Further it uses SMOTE technique to generate data form sample or Neighbour sample. | | |
| Robust hybrid data-level sampling approach (RHDLSA) [40] | This technique used multiple phases to solve the problem, in first phase its find out and remove the noise in data post that it detects majority and minority clusters by applying K-based fuzzy cluster method. In second step majority and minority cluster balancing are done. Further it used α-cut in majority class to reduce the size of data. And post that uses SMOTE to generate synthetic data and finally when all redundant data removed it uses Decision tree to balance the data. | DKFCM (Kaur et al. 2013, 2011) clustering method is being used here for separate data in majority and minority cluster and removing the noise. | Prabhjot Kaur and Anjana Gosain-2020 |

**Problem Statement**

During study, it is observed that Class Imbalance problem itself is not a big issue. It becomes a serious problem when other data impurities like class overlapping, noise, imbalance within class etc. also combine with it. Real data-sets [51,52] always contain noise, which may be of following forms like the missing values, data acquisition through electronic devices in case of medical imaging or may be because of impulse noise in case of network data etc. Keeping in view these data complexities, the class imbalance techniques must deal with the noise efficiently within the algorithm by considering other data complexities along with the class imbalance problem. From the literature review, it is noticed that there are very few techniques which tackle noise during classification of unbalanced data. Other techniques are designed based upon the assumption that either noise is not present or is cleaned before applying the algorithm. The objectives of this thesis are as follows.

- To study the existing literature on Class Imbalance Problem and identify the gaps.

- To develop an algorithm which can provide better solution to address class imbalance problem when noise is present in the dataset.
- To evaluate the propose solution and analyze the influence of imbalance related factors on the solution through performance metric designed for imbalance domains. To compare the propose solution with the existing techniques using different data-sets.

**Research Gap**

- During research review it's observed that the preprocessing of data is quite important for class imbalance problem.
- Data-level technique is effective, having bigger scope and easy to implement
- Number of available techniques in Data level hybrid approach is very less, hence it provides an opportunity to work in this area.
- Usually, imbalance class suffers with noisy and class overlapping issue in data set, so considering these two points will produce a better result.

**Research Objective**

- To Study the existing literature on Class Imbalance Problem and identify the gaps.
- To develop an algorithm which can provide solution to address class imbalance problem when noise is present in the dataset.
- To evaluate the proposed solution and analyze the influence of imbalance related factors on the solution through performance metric designed for imbalance domains. To compare the propose solution with the existing techniques using different data-sets.

**Purpose of the study**

The basic purpose of the study is to find the problem of class imbalance problem and to find a better solution of the problem.

Classification methods are very useful in solving many real-life problems. In the research literature, so many classification techniques are proposed like Decision Tree,

SVM, Neural networks etc. These classification techniques work efficiently in classifying the balanced data-sets wherein the number of instances in the classes are approximately equal. Their internal design favors the balanced data-sets. These techniques fail to detect classes when used with imbalanced data-set, because as per their internal design the results in case of unequal size of classes deviates towards the bigger class. These algorithms ignore the smaller class as noise.

In real life situations, sometimes there is a need to detect exceptional cases e.g. credit card frauds, tumor detection, fraudulent telephone calls, Shuttle system failure, text classification, oil spill detection, web spam detection, risk management, information retrieval, intrusion detection, earthquake and nuclear explosion, helicopter gear-box fault monitoring [17-18] etc. In such cases, Traditional Classification Algorithms do not work well. This problem is identified as Class Imbalance Problem (CIP). Class Imbalance problem is the classification problem wherein we are using traditional classification algorithms to classify data with unequal size classes and our objective is to identify smaller class from the data. Researchers have addressed this problem in various diversified ways and a new field of research has emerged under the name Class Imbalance Learning (CIL) and it is evolving day by day. In many papers it is referred as dealing with IDS (Imbalanced data sets) or with rare cases or dealing with skewed data sets (SDS) or skewed distributions. The smaller class in CIL is known as minority class and bigger class is known as majority class.

**Scope of the Study**

Class Imbalance Problem does not exist if the purpose is to identify majority class, it actually exists because the purpose is to identify minority class. The ratio of the number of majority to minority class data instances is called imbalance ratio. The problem becomes more risky as this ratio increases i.e. when data-set is highly imbalanced. The techniques proposed by researchers to solve the Class Imbalance Problem are majorly classified into data-level approaches (Pre-processing techniques), algorithm level approached and their hybrid [122]. In data-level approach, the researchers have tried to balance the data-sets before applying traditional classification algorithms so that results may not overwhelmed by the majority class [31,36,91,114] In algorithm level approaches, the researchers have worked upon the internal algorithm structure and tried to work upon the sensitivity of algorithm towards the majority class. These algorithms

comes under the category of cost sensitive algorithms [3,7,13] Third approach is the hybrid form, which is the combination of data-level and algorithm level approaches. The advantage of data level approaches is that researcher will work at the data level and balance the data before classification hence same classification algorithms can be used. This Chapter proposed a fast and robust hybrid data level approach based upon fuzzy logic. The proposed method can work with any level of imbalance data.

**Research Methodology and Design**

When we compares the classification performance of 20 data level (Oversampling and Undersampling) methods (listed in Table 3) with 44 imbalanced data-sets which are the variations of real data-sets at UCI (Weiss 2004). The characteristics of the data-sets are listed in Table 4.2. These methods are analyzed using KEEL software tool [7,8,13,21,67]We have used AUC (Area Under the ROC curve), F-measure and G-mean metrics to evaluate the performance of methods as these metrics are used by most of the researchers to evaluate the performance of classification algorithms. Model generated in KEEL for comparison is shown in below Figure.

**Table : Data level Methods**

| **Over-sampling Methods** | **Under-sampling Methods** |
|---|---|
| Adaptive Synthetic Sampling (ASS) | Condensed Nearest neighbor (CNN) |
| Adjusting the Direction of the Synthetic Minority class Examples (ADOMS) | Tomek Links (TL) |
| Agglomerative Hierarchical Clustering (AHC) | CNN with Tomek Links (CNNTL) |
| Random Oversampling (ROS) | Class purity maximization (CPM) |
| Synthetic Minority Oversampling Technique (SMOTE) | Neighborhood Cleaning Rule (NCL) |
| Safe Level SMOTE (SL-SMOTE) | One Sided Selection (OSS) |

| | |
|---|---|
| Borderline SMOTE (B-SMOTE) | Random Undersampling (RUS) |
| SMOTE with Edited Nearest neighbor (SMOTE_ENN) | Sampling based on Clustering (SBC) |
| SMOTE-RSB | |
| SMOTE with Tomek Links (SMOTE-TL) | |
| Selective Preprocessing of Imbalanced Data (SPIDER) | |
| SPIDER2 | |

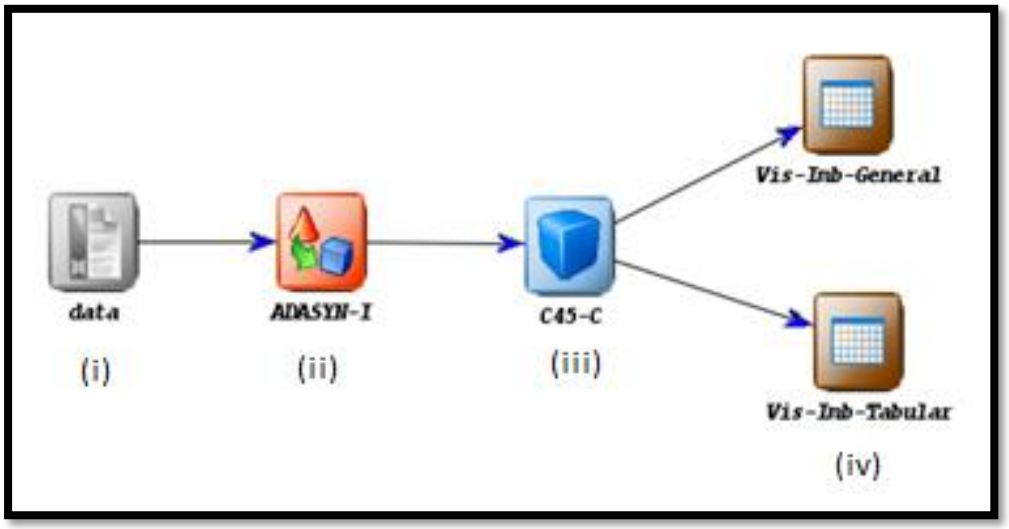



**Figure Experimental setup for comparing methods in KEEL Tool**

The setup consists of the following elements:

i. *Dataset* – In Figure 5, part (i) represents the dataset selected to study the algorithms.

ii. *Algorithm under study* – In Figure 5, part (ii) represents the algorithm that is used to balance the data-set.

iii. *Base Classifier* – In Figure 5, part (iii) represents the base Classifier. In our study, we have used decision tree (C4.5) as the base classifier. Decision tree is a very popular traditional classifier which has been extensively used by researchers to compare the methods [26]

iv. *Result* – In Figure 5, part (iv) represents the analysis in terms of confusion matrix in tabular form.

**Methods of Data Collection**

We recorded 880 confusion matrices by implementing 20 methods (Undersampling and Oversampling) on 44 UCI real data-sets. For proper analysis and representations, we have divided the 44 data-sets [78] into 6 groups based upon their class Imbalance ratio (IR) as follows: (Refer Table 4).

- Group 1: IR(1.00 – 2.00)
- Group 2: IR(2.01 – 5.00)
- Group 3: IR(5.01 – 8.00)
- Group 4: IR(8.01 – 10.00)
- Group 5: IR(10.01 – 20.00)
- Group 6: IR(>20)

We recorded the data for AUC scores for all the algorithms in each dataset separately and then calculated average values of AUC scores in each group of datasets. The results are represented in graphical form in below Figures.

**Characteristics of Real Datasets used for Simulations example**

| **Data-set** | **Group Number** | **Instances** | **Attributes** | **%Class (minority, majority)** | **IR** |
|---|---|---|---|---|---|
| glass1 | Group 1 | 214 | 9 | (35.46, 64.54) | 1.82 |
| ecoli-0_vs_1 | | 220 | 7 | (34.97, 65.03) | 1.86 |
| wisconsin | | 683 | 9 | (34.97, 65.03) | 1.86 |
| pima | | 768 | 8 | (34.84, 65.16) | 1.87 |
| iris0 | | 150 | 4 | (33.33, 66.67) | 2.00 |
| glass0 | Group 2 | 214 | 9 | (32.68, 67.32) | 2.06 |

**Research Tools**

1. Keel tool : To do data level analysis , to comparing the data set with different techniques
2. Matlab: to balance the test data, to fine different performance matrices

3. Google colab : To run the Python code to generate fine result set, graphs etc

**Proposed technique overview**

In proposed fast, robust and effective hybrid data level approach based upon fuzzy concept to handle the imbalance data. It is called fast and robust because it can handle any amount of noise in the data-set and has least time complexity compared to other methods. It is most effective in case of real time datasets because of its noise resistant nature. It can work with any level of imbalance situation. Fig. A and Fig. B show the algorithm and model of the proposed approach.

***Input:*** *Imbalance Data corrupted with noise (I-DS)*
***Output****: Balanced Dataset (B-DS)*

***Step 1****: Cluster the dataset (I-DS) into minority and majority class using DOFCM clustering approach.*

***Step 2***: *Record the Fuzzy membership values of minority and majority data points.*

***Step 3:*** *Reduce the size of majority class by removing those points whose fuzzy membership value is low.*

***Step 4:*** *Enhance the size of minority class by using modified SMOTE concept (Refer to Section II-B)*

***Step 5:*** *Combine the modified minority and majority class to generate balance dataset (B-DS).*

**Fig. A. Algorithm of Proposed method**

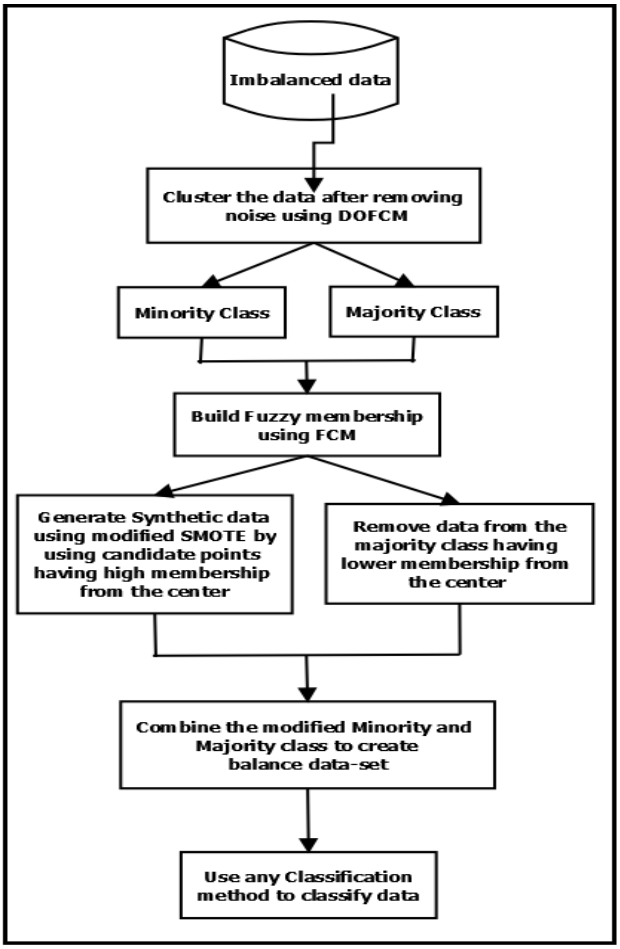

Imbalanced data
Cluster the data after removing noise using DOFCM
Minority Class
Majority Class
Build Fuzzy membership using FCM
Generate Synthetic data using modified SMOTE by using candidate points having high membership from the center
Remove data from the majority class having lower membership from the center
Combine the modified Minority and Majority class to create balance data-set
Use any Classification method to classify data

**Fig B: Model of Proposed method**

## Conclusion

In proposed work, is a data level hybrid approach has been introduced to overcome the problem of class imbalance problem in binary data set in noise environment. The data level hybrid approach is one of the best approaches considered because of its beauty to balance the data in both the minor and major classes. In this research work, a data level hybrid technique is used to balance the data with noisy environment as well. Fuzzy technique has been used to remove the noise from the row data and then to re-cluster the data. Then, the over sampling as well as under sampling have been performed based on membership matrix function. However, due to the limitation of existing fuzzy techniques, as result is changed based on the membership value selected for FCM method, The proposed technique tries to reimburse for this drawback. Moreover, Fuzzy clustering-based segmentation has also been considered as well as ensemble technique

has been used at the end to get the final outcome. The performances of these algorithms are better than their counter algorithms. The chapter Results and Simulations includes the advancement of the proposed method along with the conclusions.

Usage of density oriented fuzzy *c*-means clustering algorithm is meant to remove the noise and re-cluster the data and then hybrid techniques (oversampling and undersampling) have been used to balance the data. Clustering of numerical data forms the basis of many classification and system modeling algorithms. The purpose of clustering is to distill natural grouping of data from a large data set, producing a concise representation of a system behavior. Clustering and Density Oriented fuzzy *c*-means clustering shows that the proposed approach is characterized by higher clustering accuracy.

A comparison among the proposed technique as well as existing technique has been done in terms of performance time. The parameters chosen to execute the results are based on the performance matrix AUC (Area under the curve), F-measure and G-mean (Geometric mean) performance criteria. Moreover, the proposed technique "Fast and Robust Fuzzy Based Hybrid Data-level method to Handle Class Imbalance" produces the best result in terms of numerical values and in terms on performance time in the noised environment.

**Contributions of the Thesis**

Earlier all the techniques used to handle class imbalance at data level were either oversampling or under sampling. This research work contributes in a new way to preprocess the data by combining these two methods i.e. oversampling and undersampling keeping in mind the accuracy in preprocessing the data. This proposed technique re-clusters the data using density Metrix function and then both the methods undersampling and oversampling has been applied for balancing the data. Thus, this amalgamation of undersampling and oversampling provides more accurate results in terms of timing and value both.

## References


1. Adriaans, P. (1996). *Data mining*. Pearson Education India.

2. Ahlawat, K., Chug, A., & Singh, A. P. (2021). An Insight on the Class Imbalance Problem and Its Solutions in Big Data. In *Large-Scale Data Streaming, Processing, and Blockchain Security* (pp. 39-49). IGI Global.

3. Ai, X., Wu, J., Sheng, V. S., Zhao, P., & Cui, Z. (2015). Immune centroids oversampling method for binary classification. *Computational intelligence and neuroscience*, *2015*.

4. Akbani, R., Kwek, S., & Japkowicz, N. (2004, September). Applying support vector machines to imbalanced datasets. In *European conference on machine learning* (pp. 39-50). Springer, Berlin, Heidelberg.

5. Albrecht, J. D., & Huron, D. (2012). A statistical approach to tracing the historical development of major and minor pitch distributions, 1400-1750. *Music Perception: An Interdisciplinary Journal*, *31*(3), 223-243.

6. Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., & Herrera, F. (2011). Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic & Soft Computing*, *17*.

7. Alcalá-Fdez, J., Sánchez, L., Garcia, S., del Jesus, M. J., Ventura, S., Garrell, J. M., ... & Herrera, F. (2009). KEEL: a software tool to assess evolutionary algorithms for data mining problems. *Soft Computing*, *13*(3), 307-318.

8. Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2013). Classification with class imbalance problem. *Int. J. Advance Soft Compu. Appl*, *5*(3).

9. Al-nami, S. Y., Aljuhani, E., Althagafi, I., Abumelha, H. M., Bawazeer, T. M., Al-Solimy, A. M., ... & El-Metwaly, N. (2021). Synthesis and Characterization for new nanometer Cu (II) complexes, conformational study and molecular docking approach compatible with promising in vitro screening. *Arabian Journal for Science and Engineering*, *46*(1), 365-382.

10. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

11. Laurikkala, J. (2001, July). Improving identification of difficult small classes by balancing class distribution. In *Conference on artificial intelligence in medicine in Europe* (pp. 63-66). Springer, Berlin, Heidelberg.

12. Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., ... & Hussain, A. (2016). Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. *IEEE Access*, *4*, 7940-7957.

13. Anand, R., Mehrotra, K., Mohan, C. K., & Ranka, S. (1995). Efficient classification for multiclass problems using modular neural networks. *IEEE transactions on Neural*

*Networks*, *6*(1), 117-124.

14. Arafat, M. Y., Hoque, S., Xu, S., & Farid, D. M. (2019, August). An under-sampling method with support vectors in multi-class imbalanced data classification. In *2019 13th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)* (pp. 1-6). IEEE.

15. Bache, K., & Lichman, M. (2013). UCI Machine Learning Repository [http://archive. ics. uci. edu/ml]. University of California, School of Information and Computer Science. *Irvine, CA*, *28*.

16. Bache, K., & Lichman, M. (2013). UCI machine learning repository.

17. Bahler, D., & Navarro, L. (2000, July). Methods for combining heterogeneous sets of classifiers. In *17th Natl. Conf. on Artificial Intelligence (AAAI), Workshop on New Research Problems for Machine Learning*.

18. Barandela, R., Valdovinos, R. M., & Sánchez, J. S. (2003). New applications of ensembles of classifiers. *Pattern Analysis & Applications*, *6*(3), 245-256.

19. Barua, S., Islam, M. M., Yao, X., & Murase, K. (2012). MWMOTE--majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Transactions on knowledge and data engineering*, *26*(2), 405-425.

20. Basha, S. M., Rajput, D. S., & Vandhan, V. (2018). Impact of gradient ascent and boosting algorithm in classification. *International Journal of Intelligent Engineering and Systems (IJIES)*, *11*(1), 41-49.

21. Bashir, K., Li, T., Yohannese, C. W., & Yahaya, M. (2020). SMOTEFRIS-INFFC: Handling the challenge of borderline and noisy examples in imbalanced learning for software defect prediction. *Journal of Intelligent & Fuzzy Systems*, *38*(1), 917-933.

22. Batista, G. E., Bazzan, A. L., & Monard, M. C. (2003, December). Balancing Training Data for Automated Annotation of Keywords: a Case Study. In *WOB* (pp. 10-18).

23. Batista, G. E., Monard, M. C., & Bazzan, A. L. (2004, November). Improving rule induction precision for automated annotation by balancing skewed data sets. In *International Symposium on Knowledge Exploration in Life Science Informatics* (pp. 20-32). Springer, Berlin, Heidelberg.

24. Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, *6*(1), 20-29.

25. Kubat, M., & Matwin, S. (1997, July). Addressing the curse of imbalanced training sets: one-sided selection. In *Icml* (Vol. 97, No. 1, p. 179).

26. Błaszczyński, J., Deckert, M., Stefanowski, J., & Wilk, S. (2010, June). Integrating selective pre-processing of imbalanced data with ivotes ensemble. In *International conference on rough sets and current trends in computing* (pp. 148-157). Springer, Berlin, Heidelberg.

27. Garcia, S., & Herrera, F. (2008). An Extension on" Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons. *Journal of machine learning research*, *9*(12).

28. Mani, I., & Zhang, I. (2003, August). kNN approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of workshop on learning from imbalanced datasets* (Vol. 126, pp. 1-7). ICML

29. Breiman, L. (1999). Pasting small votes for classification in large databases and on-line. *Machine learning*, *36*(1), 85-103.

30. Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2009, April). Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 475-482). Springer, Berlin, Heidelberg.

31. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2004). A review on Random over sampling in class imbalance problem. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(4), 463-484.

32. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, *21*(9), 1263-1284

33. He, H., Bai, Y., Garcia, E. A., & Li, S. (2008, June). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* (pp. 1322-1328). IEEE

34. Hu, S., Liang, Y., Ma, L., & He, Y. (2009, October). MSMOTE: improving classification performance when training data is imbalanced. In *2009 second international workshop on computer science and engineering* (Vol. 2, pp. 13-17). IEEE

35. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

36. Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, *6*(1), 1-6.

37. Liang, X. W., Jiang, A. P., Li, T., Xue, Y. Y., & Wang, G. T. (2020). LR-SMOTE—An improved unbalanced data set oversampling based on K-means and SVM. *Knowledge-Based Systems*, *196*, 105845.

38. Ramentol, E., Caballero, Y., Bello, R., & Herrera, F. (2012). SMOTE-RSB*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced

data-sets using SMOTE and rough sets theory. *Knowledge and information systems*, *33*(2), 245-265.

39. Al-Rifaie, M. M., & Alhakbani, H. A. (2016, July). Handling class imbalance in direct marketing dataset using a hybrid data and algorithmic level solutions. In *2016 SAI Computing Conference (SAI)* (pp. 446-451). IEEE.

40. Kaur, P., & Gosain, A. (2020). Robust hybrid data-level sampling approach to handle imbalanced data during classification. *Soft Computing*, *24*(20), 15715-15732.

41. Chi, Z., Yan, H., & Pham, T. (1996). *Fuzzy algorithms: with applications to image processing and pattern recognition* (Vol. 10). World Scientific.